# Norris outlines NIH Big Data initiative

*By Richard Sloane*

New efforts are underway to manage the huge amounts of scientific data being produced by researchers throughout the National Institutes of Health, according to NIH Chief Information Officer Andrea Norris, who gave a talk March 29 at NIEHS about the NIH Big Data to Knowledge (BD2K) *(http://nexus.od.nih.gov/all/2013/02/28/big-data-and-the-biomedical-workforce-nih-wants-your-input/)* initiative.

*Norris was candid about the current state of information technology at NIH. She told the capacity audience. "It'll take us a few years to catch up." (Photo courtesy of Steve McCaw)*

Norris explained that BD2K aims to create improved data and software sharing policies, catalogs of research data, and the development of data and metadata quality standards. She said she expects to see a significant long-term investment by NIH for accelerated software development and enhanced training through new biomedical big data centers of excellence.

According to Norris, the BD2K initiative would create an advanced computing environment called InfrastructurePlus, which would ultimately modernize the NIH network to meet data handling requirements through a much more robust network. Ideally, InfrastructurePlus would advance high-performance computing, and agile hosting and storage approaches for different data domains.

"Fundamental change in the way we gather and use massive amounts of data is overdue," said Norris. The aging NIH computer network currently runs at 80 to 90 percent capacity during peak utilization, far higher than the desirable 30 to 40 percent. According to Norris, meeting the challenge of managing growing amounts of data (see text box) involves more sophisticated technology, a dedicated research network, better harmonization tools, and even cultural evolution.

*Employees from every division of NIEHS attended Norris' presentation, including NIH Human Resources Specialist Angela Davis, left, and NIEHS Employee Services Manager Ed Kang. (Photo courtesy of Steve McCaw)*

### Issues and opportunities

"Big data is changing dramatically how we do science," Norris said. "Accessing these massive pools of data will most likely require new skill sets by scientists."

More scientists are increasingly using pooled data, instead of working with only their own. In many circles of science, teams of researchers are leveraging large, and even massive, amounts of data (see story).

This new kind of shared data approach challenges the culture of scientific research, Norris said, because it will require the community to recognize the value of generating good data, and allowing access to that data. The research culture at NIH will need to change, to respond effectively to new developments in an ever-changing technology landscape.

### Many questions remain

Bioinformatics, genetics, and genomics studies produce and consume massive amounts of data. Yet many questions arise. How will this data be stored? How will it be accessed? Who will be in control? What are the hardware and software challenges to facilitate big data management? How will data be shared? How can data quality be assured and adaptable to the needs of science? How long should data be stored?

"We are learning," Norris explained. "In five years we'll look back and realize how naïve we were on some of these approaches. Whatever we're doing today, we expect to adapt, mature, and evolve over time."

(Richard Sloane is an employee services specialist with the NIEHS Office of Management.)

*Also part of the audience were, left to right, NTP Deputy Division Director for Policy Mary Wolfe, Ph.D.; NTP Deputy Director of the Office of Health Assessment and Translation Andrew Rooney Ph.D.; and NIEHS Informationist Stephanie Holmgren. (Photo courtesy of Steve McCaw)*

*Senior Associate Scientist Dmitry Gordenin, Ph.D., was one of many researchers who expressed their interest in better tools for harnessing big data, and open access to government-funded research. (Photo courtesy of Steve McCaw)*

*Despite the seriousness of the current situation, Norris could also speak to the lighter side of NIH culture. Shown above enjoying the comic relief, left to right, were Kang; NIEHS Project Officer Beth Ragan; Program Administrator David Balshaw, Ph.D.; Health Science Administrator Symma Finn, Ph.D.; and Grants Management Specialist Molly Puente, Ph.D. (Photo courtesy of Steve McCaw)*

*Birnbaum joined Norris during the question and answer session. Birnbaum described effective data management as critical for advancing strategic plan initiatives. (Photo courtesy of Steve McCaw)*

## Big Data means very big numbers

Big Data is measured in terabytes, or trillions of bytes, and petabytes, or quadrillions of bytes, but according to some experts, within a decade, even those numbers may be inadequate.

According to estimates by Eric Schmidt, Google's former chief executive officer, the world creates 5 exabytes, or quintillions of bytes, of data every two days — roughly the same amount of data created between the dawn of civilization and 2003.

It's estimated that NIH generates 4 petabytes of data each day.

---

The Environmental Factor is produced monthly by the National Institute of Environmental Health Sciences (NIEHS) *(http://www.niehs.nih.gov/)*, Office of Communications and Public Liaison. The content is not copyrighted, and it can be reprinted without permission. If you use parts of Environmental Factor in your publication, we ask that you provide us with a copy for our records. We welcome your comments and suggestions. *(brus kec@niehs.nih.gov)*